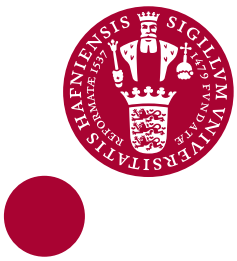

Fast and powerful hashing using tabulation

Thorup, Mikkel

Published in:
44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)

DOI:
10.4230/LIPIcs.ICALP.2017.4

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Fast and Powerful Hashing Using Tabulation*†

Mikkel Thorup

Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
mikkel2thorup@gmail.com

Abstract

Randomized algorithms are often enjoyed for their simplicity, but the hash functions employed to yield the desired probabilistic guarantees are often too complicated to be practical. Here we survey recent results on how simple hashing schemes based on tabulation provide unexpectedly strong guarantees.

Simple tabulation hashing dates back to Zobrist [1970]. Keys are viewed as consisting of c characters and we have precomputed character tables $h_1, ..., h_q$ mapping characters to random hash values. A key $x = (x_1, ..., x_c)$ is hashed to $h_1[x_1] \oplus h_2[x_2] ..... \oplus h_c[x_c]$. This schemes is very fast with character tables in cache. While simple tabulation is not even 4-independent, it does provide many of the guarantees that are normally obtained via higher independence, e.g., linear probing and Cuckoo hashing.

Next we consider *twisted tabulation* where one character is "twisted" with some simple operations. The resulting hash function has powerful distributional properties: Chernoff-Hoeffding type tail bounds and a very small bias for min-wise hashing.

Finally, we consider *double tabulation* where we compose two simple tabulation functions, applying one to the output of the other, and show that this yields very high independence in the classic framework of Carter and Wegman [1977]. In fact, w.h.p., for a given set of size proportional to that of the space consumed, double tabulation gives fully-random hashing.

While these tabulation schemes are all easy to implement and use, their analysis is not.

This keynote talk surveys results from the papers in the reference list.

1998 ACM Subject Classification E.1 [Data Structures] Tables, E.2 [Data Storage Representations] Hash Table Representations, F.2.2 [Analysis of Algorithms and Problem Complexity] Nonnumerical Algorithms and Problems – Sorting and Searching, H.3 [Information Storage and Retrieval] Information Search and Retrieval – Search Process

Keywords and phrases Hashing, Randomized Algorithms

Digital Object Identifier 10.4230/LIPIcs.ICALP.2017.4

Category Invited Talk

* Research partly supported by Advanced Grant DFF-0602-02499B from the Danish Council for Independent Research.

† A similar talk abstract also appears in *Proceedings of the 36th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, pages 1:1–1:2, 2016 – http://dx.doi.org/10.4230/LIPIcs.FSTTCS.2016.1 –, and in *Proceeding of the 14th ACM/SIGEVO Conference on Foundations of Genetic Algorithms (FOGA), page 1–1, 2017.*

References

1. Tobias Christiani, Rasmus Pagh, and Mikkel Thorup. From independence to expansion and back again. In *Proceedings of the 47th ACM Symposium on Theory of Computing (STOC)*, pages 813–820, 2015.
2. Søren Dahlgaard, Mathias Bæk Tejs Knudsen, Eva Rotenberg, and Mikkel Thorup. The power of two choices with simple tabulation. In *Proceedings of the 27th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1631–1642, 2016.
3. Søren Dahlgaard and Mikkel Thorup. Approximately minwise independence with twisted tabulation. In *Proc. 14th Scandinavian Workshop on Algorithm Theory (SWAT)*, pages 134–145, 2014.
4. Søren Dahlgaard, Mathias Bæk Tejs Knudsen, Eva Rotenberg, and Mikkel Thorup. Hashing for statistics over k-partitions. In *Proceedings of the 56th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1292–1310, 2015.
5. Mihai Pătraşcu and Mikkel Thorup. The power of simple tabulation-based hashing. *Journal of the ACM*, 59(3):Article 14, 2012. Announced at STOC'11.
6. Mihai Pătraşcu and Mikkel Thorup. Twisted tabulation hashing. In *Proc. 24th ACM/SIAM Symposium on Discrete Algorithms (SODA)*, pages 209–228, 2013.
7. Mikkel Thorup. Simple tabulation, fast expanders, double tabulation, and high independence. In *Proc. 54th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 90–99, 2013.